The Phrasal Lexicon in Multilingual Text Generation

Cornelia M. Verspoor
Intelligenesis Corporation[0]
50 Broadway, Suite 1205
New York, NY 10004
karin@intelligenesis.net

Vicente Uceda
Logos Corporation[0]
100 Enterprise Drive, Suite 501
Rockaway, NJ 07866
uceda@logos-usa.com

Paper ID Code: get id
Submission Type: General Session
Topic Areas or Theme ID: multilingual natural language generation
Word Count: 3150
Under consideration for other conferences (specify)? none

Abstract

We introduce the ML-Peba system, a system which generates textual descriptions of animals in both English and Spanish from an abstract knowledge base. The system relies upon a *phrasal lexicon* for linguistic realization. The paper discusses the pros and cons of this approach for multilingual generation, suggesting that its use enables a more language-independent knowledge representation than most generation approaches.

[0] The work reported in this paper was completed while both authors were employed at the Microsoft Research Institute, Macquarie University, Sydney, Australia. It was supported with a Macquarie University Research Grant.

1 Introduction

Natural language generation (NLG) systems aim to produce texts from an abstract representation of information. One fundamental premise of this work is that knowledge representation in this context is not structured according to realization in a specific language. The possibility of multilingual generation systems follows from this premise: the knowledge source is language-neutral, supporting linguistic expression in potentially any language. Only the mapping from that source to a text will be influenced by the structure of individual languages.

There is, however, considerable uncertainty as to what it means for a representation to be "language-neutral" in the generation context, given that the system must find a correspondence between the representation and a linguistic realization and that the design of the knowledge representation is often guided by that requirement. This issue echoes the question in interlingua-based machine translation systems as to what the appropriate structure of an interlingua is.

One of the most prevalent bases for knowledge representation utilized in NLG work is the Upper Model (Bateman, 1990), which provides an ontology of basic linguistic distinctions, to which domain knowledge is anchored. In principle, the ontology provided by the Upper Model is language-independent; it does not embody specific assumptions deriving from any particular language. However, in practice, structuring domain knowledge according to a linguistically-motivated ontology can require a division of concepts which is unnatural to the domain (Stede and Grote, 1995), and can require ongoing refinement of the knowledge representation due to differences in how languages partition phenomena (Matiasek and Trost, 1995).

A truly language-independent conceptual representation need not even be constructed specifically for the NLP task; it could, for example, be a relational database previously constructed for another task. Utilizing previously existing resources makes sense in terms of time and money.

As a result of these concerns, we believe that multilingual NLG systems must be architected to keep domain knowledge and linguistic knowledge distinct. The surface realization components of an NLG system must take on the responsibility of structuring domain knowledge in language-specific ways for output. We suggest that the use of a *phrasal lexicon* enables knowledge representation to be largely language-neutral, by providing a direct mapping from complex domain knowledge to linguistic realization. It then follows that the phrasal lexicon provides a hightly suitable framework for multilingual generation. We will show this through the extension of a monolingual (English) NLG system, the Peba-ii system (Milosavljevic et al., 1996) with an additional output language (Spanish).

2 The Phrasal Lexicon

In a phrasal lexicon, lexical entries associate complex concepts with linguistically complex realizations. In contrast to the traditional NLG lexicon in which a concept maps to a single word, requiring complex reasoning in the surface realization component to combine words together into phrases, the complexity under a phrasal lexicon approach is embedded in phrasal elements placed directly in the lexicon. The surface realization component of the generation system only controls the combination of high-level linguistic units, such as the juxtaposition of a noun phrase and a verb phrase to form a sentence.

Milosavljevic et al. (1996) argue that since knowledge representation for a particular domain often uses complex concepts, the linguistic elements to which they correspond should be equally complex. So, for example, the Peba-ii lexicon associates the knowledge base concept `eats-ants-termites-earthworms` with the phrase "is a carnivore and eats ants, termites and earthworms".[1]. Milosavljevic et al. (1996) state,

> The use of phrasal lexical items of this kind has two specific advantages: **Reuse and Efficiency**. If we repeatedly realize a semantic element in the same way, it is better to remember this and avoid rebuilding the surface form each time.

Such phrasal units must exist in the lexicon regardless, due to the existence of idioms and other non-compositional linguistic material (Verspoor, 1997).

We propose an additional advantage of the phrasal lexicon: it allows language-specific variation to be localized in the lexicon, relieving the domain knowledge of any requirement of compatibility with a linguistic ontology and thereby enabling that knowledge to easily support multilingual applications. We will show this through our extension of Peba-ii system to produce output in Spanish as well as in English, **without any change** in the underlying knowledge representation.

3 The Peba-II system

We take as our starting point the Peba-ii system. This is a web-based NLG system which generates English-language description and comparison hypertexts on the subject of animals. The architecture of the system is described in detail in Milosavljevic et al. (1996).

Content in Peba-ii is represented in a taxonomic knowledge base structured according to the Linnæan classification. Neither a linguistic ontology nor linguistic considerations were utilized in the construction of this knowledge base, although it was constructed specifically for use in

[1] The phrasal elements in the Peba-ii system are hand-constructed, but the use of automatic techniques for constructing a phrasal lexicon is being explored in another project at the Microsoft Research Institute, the Power system (Dale et al., 1998).

```
(hasprop Echidna
  (linnaean-classification Family))
(distinguishing-characteristic Echidna Monotreme
  (body-covering sharp-spines))
(hasprop Echidna
  (geography found-Australia))
(hasprop Echidna
  (social-living-status lives-by-itself))
(hasprop Echidna
  (diet eats-ants-termites-earthworms))
(hasprop Echidna
  (length (quantity (lower-limit (unit cm) (number 35))
                    (upper-limit (unit cm) (number 60)))))
(hasprop Echidna
  (weight (quantity (lower-limit (unit kg) (number 2))
                    (upper-limit (unit kg) (number 7)))))
```

Figure 1: A portion of the PEBA-II knowledge base

```
Identify:
        Name Entity
        Clarificatory Comparison
        List Subtypes
        Describe Properties

Name Entity:
        Primary Name
        Secondary Name (if available)
        Supertype
        Relationship of the supertype to the entity
        Distinguishing characteristic (if there is one) of this
           entity relative to other subtypes of the supertype
```

Figure 2: Two discourse plans from PEBA-II

this system. An example of PEBA-II's knowledge base appears in Figure 1.

In general, each animal property is a typed concept in the knowledge base which has a directly associated expression as a verb phrase listed in the PEBA-II phrasal lexicon. The concepts for animal names and classifications are associated with noun phrases.

A surface realization component then controls the combination of these “chunks” to form sentences. The PEBA-II system has no explicit grammatical rules governing the combination of the elements in the phrasal lexicon. Instead, sentence templates are defined for each component of a discourse plan. They are essentially sentences with “holes” – gaps filled in by entries from the phrasal lexicon. The templates rely on the basic grammatical division introduced above.

Two typical Peba-ii discourse plans are shown in Figure 2. The *Identify* plan structures information at the level of a paragraph, while the *Name Entity* plan structures information at the level of an individual sentence. The abstract content labels in those discourse plans are replaced with specific concepts as the system processes a generation request. A specific *Name Entity* plan is then realised using a template which results in the construction of a sentence like the following:

> *The <Echidna primary-name>, also known as the <Echidna secondary-name>, is a type of <Echidna supertype> which <distinguishing-property of Echidna>.*

The phrasal lexicon is consulted for the linguistic realization of the concepts, which results in the following text:

> *The Echidna, also known as the spiny anteater, is a type of Monotreme which is covered in stiff, sharp spines mixed with long, coarse hairs.*

The fluency of the text constructed from this template depends on the consistent grammatical divisions, and the quality of the entries in the phrasal lexicon. The generation system does not have to reason about valid combinations of phrases.

4 Extending Peba-II for multilinguality

4.1 ML-Peba: Methodology

In extending the Peba-ii system to ML-Peba (Multilingual Peba)[2], supporting generation of texts in Spanish, we took a *minimal changes* approach to test the ability of the underlying

[2] This system is available for testing at `http://www.mri.mq.edu.au/~peba/MLPeba/system.html`

knowledge base to support Spanish as well as English.

Our methodology was as follows:

1. Translate each (phrasal) lexical entry into Spanish, adding gender and syntactic number specifications and agreement verification.
2. Restructure the templates in the sentence planner to support both Spanish and English realizations for all abstract sentence types in the discourse plans, without changing the general sentence plans to accommodate Spanish.
3. Maintain the knowledge base unchanged.

After these steps were followed, a full set of descriptions and comparisons was generated in Spanish with ML-Peba and evaluated for errors and fluency.

4.2 ML-Peba: Results

The transfer of the PEBA-II system to the task of generation in multiple target languages was relatively straightforward, although time-consuming due to the need for translation of the large phrasal lexicon. The texts which the extended system, ML-Peba, generates in Spanish display no grammatical errors and are surprisingly fluent.

Figure 3 is an example of an English text generated to ML-Peba; the corresponding Spanish text is in Figure 4. Both texts are judged to be fully grammatical and quite fluent.

The African Porcupine is a type of Rodent that has long sharp spines, up to 50cm long, which cover its whole back and can be raised by muscles under the skin.

Although it is similar in appearance to the Echidna, it is not closely related. The Echidna, also known as the spiny Anteater, is a type of Monotreme that is covered in stiff, sharp spines mixed with long, coarse hairs. Like the Echidna, the African Porcupine has a browny black coat and paler-coloured spines.

Figure 3: An English description generated by ML-Peba

El puercoespín africano es un tipo de Roedor que posee púas agudas de hasta 50 cm de longitud, que cubren toda su espalda y que levanta mediante músculos que posee debajo de la piel.

Aunque es similar en apariencia al Equidna, no están estrechamente relacionados. El Equidna, también conocido como el Oso hormiguero con púas, es un tipo de Monotrema que está cubierto de espinas duras y afiladas y de pelo duro y largo. Como el Equidna, el puercoespín africano tiene pelo marrón oscuro y espinas de color más pálido.

Figure 4: A Spanish description generated by ML-Peba

4.2.1 The utility of the phrasal lexicon

The use of a phrasal lexicon allows the structure of generated texts to be adapted to the expression requirements of each output language. This is so because the system depends on only a few syntactic rules. We can structure a phrase in a particular language according to the preferred manner of expression in that language. For example, the constituent order within a verb phrase expressing a property need not be parallel in English and Spanish.

Consider the ML-Peba lexical entry in Figure 5. The order of the constituents in the English phrase differs from the preferred order in Spanish. The Spanish phrasal structure has therefore been adapted in the lexicon to express the ideal. Thus, the adverb *often* has been replaced by a verb of frequency that better expresses the concept in Spanish and the present verb form *has* has been replaced by the infinitive form *tener* (to have). So rather than translating the English lexicalization more literally as *tiene una cola larga, a menudo prensil (agarradera)*, we obtain the more fluent phrase shown in Figure 5.

```
(lex long-prehensile-tail
  (language
   (english ((orth "has a long, often prehensile (grasping)
                    tail")
             (syn ((cat vp) (agr ((number singular))))))
   (spanish ((orth "suele tener una cola prensil (agarradera)
                    y larga")
             (syn ((cat vp) (agr ((number singular))))))))
```

Figure 5: A multilingual lexical entry in ML-Peba

```
(lex powerful-forelimbs-are-spade-like-tipped-strong-
              claws-used-burrow-through-soil
     (language
      (english ((orth "has powerful forelimbs which are spade-like
                       and tipped with strong claws used to
                       burrow through soil")
                (syn ((cat vp) (agr ((number singular))))))
      (spanish ((orth "tiene robustas patas anteriores en forma de
                       azada y fuertes garras que utiliza para
                       hacer tuneles")
                (syn ((cat vp) (agr ((number singular))))))))
```

Figure 6: Another multilingual lexical entry in ML-Peba

It would be difficult to accommodate such differences if we were constructing the sentences from their basic components. Under the phrasal lexicon approach, however, they are directly recorded in the lexicon.

Another example is found in the lexical entry in Figure 6. Here there are significant differences between the lexicalizations represented for the two output languages. The most obvious is the free translation of *to burrow through soil* as *para hacer túneles (for making tunnels)* instead of word by word as *para excavar através del suelo*. The structure of the two phrases also differs substantially, in that the full conjunctive verb phrase *are spade-like and tipped with strong claws used to burrow through soil* modifies *forelimbs* in the English phrase, while in Spanish the conjunction is of two noun phrases. Lower-level grammatical differences include the transformation of the English relative clause *which are ...* into a reduced relative clause in Spanish and the adaptation of the English passive structure *claws used to ...* in an active structure *garras que utiliza para (claws which are used for ...)*.

So variations in the best manner of expression of a concept in different languages can be directly accommodated in the phrasal lexicon approach. A human translator is allowed to decide how a concept can best be expressed in a target language, without regard to structural or indeed semantic constraints in the English original. The only constraint in ML-Peba stems from the grammatical constraints on concept realizations. As long as a naming concept is realized as a noun phrase and a property concept is realized as a verb phrase, any within-phrase differences in expression can be accommodated. This follows directly from the fact that the system does not have to reason about the internal composition of the phrases.

4.2.2 Agreement and Surface realization

In the original English system, the article "The" was added in front of nouns directly in the sentence plans, as is apparent in the Name Entity plan introduced above. For English this is possible because there is only one form of the definite article. In Spanish, however, there are four forms of the definite article (*el, la, los, las*), which vary depending on the number and gender of the noun introduced by the article. Spanish requires agreement verification between the determiner and the noun. We therefore created lexical entries for the various forms of the definite article, added a feature for gender in the syntactic representation in the phrasal lexicon, and enforced syntactic agreement between determiners and nouns, and between subject noun phrases and main verb phrases. This agreement verification in the sentence planner was the only major change required of the original Peba-ii implementation.

4.2.3 Stylistics

The structure dictated by the sentence planner for the English sentences, that is, naming Noun Phrase followed by property Verb Phrase, was kept in place for the Spanish sentences. This constraint did not impact on the grammaticality of the Spanish sentences produced by ML-Peba because English and Spanish have a parallel structure at the sentence level, but it did impact on the naturalness of those sentences.

For example, the sentence corresponding to (1a) in Spanish as generated by ML-Peba is in (1b), while the more natural translation is as appears in (1c).

(1) a. The bird has feathers and scales on its legs

b. El pájaro tiene plumas y escamas en las patas. (OK)

c. Las patas de los pájaros están recubiertas de plumas y escamas. (preferable – *The legs of the bird are covered with feathers and scales*)

The latter structure cannot be generated in the current implementation of ML-Peba. This is because we restricted the plans to general structures valid for both target languages.

4.2.4 Ambiguity

In this system, ambiguity was not a problem because of the high-level divisions made and the large granularity of the lexicalizations in the phrasal lexicon; differences in expression could be localized inside of the complex property verb phrases. This of course would not be true in a system for which a single concept mapped to a single word; there will often be cases in which a concept which maps to a single word in one language can potentially map to multiple words in another (due to differences in conceptual divisions), and the system would have to be given some facility for choosing the appopriate lexicalization for a given concept in context.

For instance, ambiguity stems from the English verb “to be”; in Spanish it corresponds to two different verbs, “ser” and “estar”. For example:

(2) a. Although it **is** similar in appearance to the Echidna, it **is** not closely related

b. Aunque **es** similar en apariencia al Equidna, ambos no **están** estrechamente relacionados

Because the verb appropriate to particular concepts (e.g. `similar-appearance` nad `not-related`) is recorded in the lexical entry for those concepts, the system does not need to choose between them; indeed, it does not even need to understand that the verb "to be" is involved in the expression of those concepts at all.

We also avoid the problems that the generation of passive clauses. In Spanish there are two different passives (the "normal" passive and the "impersonal" passive), and the use of a phrasal lexicon means that we do not need to introduce mechanisms into the generation system for determining when to use one or the other. The examples below show the complexity of this variation.

(3) a. The problem can be solved by studying it carefully.

b. Se puede solucionar el problema examinádolo detenidamente.

c. El problema puede ser solucionado examinádolo detenidamente.

(4) a. The Colossus was invented by Alan Turing.

b. El Coloso fue inventado por Alan Turing.

c. *El Coloso se inventó por Alan Turing.

(5) a. The Colossus was invented in 1946.

b. El Colosus se inventó en 1946.

c. El Colosus fue inventado en 1946.

In (3) and (5), both passive constructions are possible, while in (4) the impersonal passive is not possible. This is because the impersonal passive can never be used when there is an agent specified. Making this decision using a different surface realization technique would require some representation of agentivity and a corresponding reasoning process.

5 Discussion

The phrasal lexicon is a highly practical approach for multilingual generation because linguistic differences between languages can be localized within the phrasal realizations in the lexica of different languages. So, while the concept `eats-ants-termites-earthworms` may best be expressed in dramatically different ways in English and Spanish, as long as both of these languages express property concepts consistently as verb phrases the differences in expression can simply be reflected in the realizations listed in the phrasal lexicons for the two languages.

The examples in Section 4.2.3 show that the strict syntactic division assumed in ML-Peba for specific types of information represented in the knowledge base can lead to unnatural sentences in Spanish; accommodation of the natural Spanish structure in more cases would require additional structure for properties – more fine-grained typing – in the PEBA-II knowledge representation. The *colouring* properties in the knowledge base, for instance, sometimes mention a color applied to a whole animal (the Woolly Oppossum "is a golden colour), and sometimes are more specific (the Black-shouldered Opossum "has a black ring around its neck and shoulders which extends down its back"). This difference impacts on the most natural expression of the *colouring* properties in Spanish (in the first example, the structure in Spanish should be *The color of the Woolly Oppossum is golden* while in the second example the structure should be as in English). The values of the *colouring* properties would have to be subdivided into two types to accommodate two distinct surface realizations. While this indicates that the requirements of linguistic expression can impact on the knowledge representation in this approach just as in any other NLG approach, such needed refinement of the representation is unlikely to arise often, and in this case mainly highlights the fact that the knowledge base in Peba-ii was implemented specifically for English generation, rather than for any non-linguistic application (which would probably also have required finer-grained divisions of concepts).

6 Conclusions

It is clear from the good results we have achieved using a phrasal lexicon and a simple surface realization methodology in the ML-Peba system that it is possible to support multilingual generation using a single knowledge representation. The use of high-level conceptual divisions and complex corresponding lexicalizations allows much of the linguistic variation between English and Spanish to be localized within the lexical entries.

This conclusion, however, is only valid more generally to the extent that two languages are parallel in structure. Had word order in Spanish been significantly different from that in English, the specific approach outlined in this paper would not have been effective. Support of the two languages in that case would have required at least a distinct surface realization component for each language, which would reason about the information in the knowledge representation in language-specific ways. It may also have required more changes to the underlying knowledge representation; however these changes would certainly be fewer than those required in an approach where domain knowledge is tied directly to linguistic knowledge.

We have seen that despite the high degree of similarity between English and Spanish at the representational level embodied by ML-Peba, there are certain stylistic problems that cannot be resolved given the structure of the knowledge representation in the system. These stylistic problems point to the weakness of the ML-Peba system as implemented – it relies on conceptual and structural similarities between the target languages. Yet the successful and straightforward extension of the PEBA-II system to include grammatical and fluent output in Spanish, without the addition of any reasoning about sentence construction specific to Spanish, also indicate the strengths of the phrasal lexicon approach in general: complex concepts do not need to be broken down in highly language-specific ways, and reasoning about the combination of atomic concepts for specific linguistic realizations can be significantly reduced by using larger-grained basic concepts. Knowledge representation does not need to be tied to a linguistic ontology to enable multilingual natural language generation.

References

John Bateman. 1990. Upper modeling: Organizing knowledge for natural language processing. In Kathleen McKeown, Johanna Moore, and Sergei Nirenburg, editors, *Proceedings of the Fifth International Workshop on Natural Language Generation*, Dawson, PA, June 3-6.

Robert Dale, Stephen J Green, Maria Milosavljevic, Cécile Paris, Cornelia Verspoor, and Sandra Williams. 1998. The realities of generating natural language from databases. In *Proceedings of the Australian Joint Conference on Artificial Intelligence.*

Johannes Matiasek and Harold Trost. 1995. Requirements on linguistic knowledge sources for multilingual generation. In *Working Notes of the IJCAI workshop on Multilingual Text Generation*, pages 102–109, Montreal, August 20-21. International Joint Conference on Artificial Intelligence.

Maria Milosavljevic, Adrian Tulloch, and Robert Dale. 1996. Text generation in a dynamic hypertext environment. In *Proceedings of the 19th Australasian Computer Science Conference*, Melbourne, Australia.

Manfred Stede and Brigitte Grote. 1995. The lexicon: Bridge between language-neutral and language-specific representations. In *Working Notes of the IJCAI workshop on Multilingual Text Generation*, pages 129–135, Montreal, August 20-21. International Joint Conference on Artificial Intelligence.

Cornelia Maria Verspoor. 1997. *Contextually-Dependent Lexical Semantics.* Ph.D. thesis, Centre for Cognitive Science, University of Edinburgh.